Модел. и анализ информ. систем. Т. **20**, № **1** (2013) 107–115

УДК 519.987

Несмещенная оценка энтропии для бинарных потоков

Тимофеев Е. А.[1]

Ярославский государственный университет им. П.Г. Демидова
150000 Россия, г. Ярославль, ул. Советская, 14

e-mail: timofeevea@gmail.com

получена 12 января 2013

Ключевые слова: энтропия, непараметрическая оценка, шар, мера Бернулли

Предлагается новый класс метрик на пространстве правосторонних бесконечных последовательностей над бинарным алфавитом. Показано, что параметры, определяющие этот класс метрик, можно выбрать так, что смещение оценки энтропии будет $O(n^{-c})$, где n – число заданных последовательностей, c – некоторая константа.

Введение

Рассматривается задача нахождения (оценивания) энтропии меры по заданным $n+1$ точкам из пространства Ω правосторонних последовательностей над конечным алфавитом $\mathcal{A} = \{0, 1\}$, которые будем считать независимыми случайными величинами, одинаково распределенными по этой мере μ.

Для нахождения обычно используются непараметрические оценки. Будем применять метод "ближайшей точки" и непараметрические оценки, предложенные в [5]. Эти оценки строятся по метрике и некоторому вспомогательному целочисленному параметру k. Для этих оценок показано, что дисперсия убывает со степенным порядком по n при общих и естественных ограничениях на меру и метрику.

Наибольшую трудность (как и для большинства непараметрических оценок) представляет нахождение смещения оценки, которое является основным препятствием для определения точности оценки.

В [5] показано, что, если метрика такая, что усредненная по пространству мера шара является гладкой функцией от радиуса, то оценка будет несмещенной. Однако в пространстве последовательностей очень трудно найти метрику, для которой будет выполнено это условие.

[1]Работа выполнена при поддержке гранта Правительства РФ по постановлению №220, договор 11.G34.31.0053.

Более того, неудачный выбор метрики может привести к новым трудностям. Так, для простейшей метрики, которая зависит только от первого несовпадения символов, в [6] показано, что смещение будет периодической функцией от $\log n$ для мер Бернулли с рационально соизмеримыми логарифмами вероятностей. Для несоизмеримых значений оценка является асимптотически несмещенной.

Таким образом, неудачный выбор метрики приводит к очень сложному смещению, которое является разрывной функцией от параметров меры.

Подчеркнем, что для бинарного алфавита это смещение очень мало (порядка 10^{-6} [4]), поэтому его невозможно заметить в вычислительном эксперименте и нужна теоретическая оценка смещения хотя бы в некотором классе метрик.

Для устранения смещения предлагается следующий подход. Выбирается достаточно широкий класс метрик, зависящих от параметров. Построенные оценки оптимизируются по параметрам метрик так, чтобы уменьшить смещение.

Итак, нужный класс метрик должен удовлетворять следующим условиям.

1. Любая метрика из этого класса эквивалентна простейшей метрике, которая зависит только от первого несовпадения символов.

2. Зависимость оценки от параметров должна быть достаточно простой, чтобы легко решалась задача оптимизации по этим параметрам.

3. Усредненная обратная функция к мере шара должна быть как можно более гладкой.

4. Усредненную обратную функцию к мере шара можно найти для бернуллиевских мер.

5. Все расстояния до ближайших точек находятся с трудоемкостью $O(n \log n)$.

В статье будет описан такой класс метрик и будет показано, что выполнены вышеописанные условия.

1. Метрики в пространстве последовательностей

Пусть $\boldsymbol{x} = (x_1, x_2, \dots)$ и $\boldsymbol{y} = (y_1, y_2, \dots)$ – две точки из $\Omega = \mathcal{A}^{\mathbb{N}}$. Через $a\boldsymbol{x}$ будем обозначать последовательность $(a, x_1, x_2, \dots)$. Рассматриваемый класс метрик определяется следующим образом:

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = e^{-\alpha(\boldsymbol{x}, \boldsymbol{y})}, \tag{1}$$

где функция $\alpha(\boldsymbol{x}, \boldsymbol{y})$ задается рекуррентным образом

$$\alpha(a\boldsymbol{x}, b\boldsymbol{y}) = \begin{cases} \alpha(\boldsymbol{x}, \boldsymbol{y}) + 1, & a = b; \\ \lambda(\alpha(\boldsymbol{x}, \boldsymbol{y})), & a \neq b. \end{cases} \tag{2}$$

Здесь вспомогательная функция $\lambda(t)$ является неубывающей и

$$\lambda(0) = 0, \quad \lambda(t) \leq 1.$$

Функцию $\lambda(t) = \lambda_{0,0}(t)$ будем выбирать из следующего параметрического семейства:

$$\begin{cases} \lambda_{i,j}(t+1) = (1-\beta_{i,j})\lambda_{i+1,2j}(t) + \beta_{i,j}, & t \geq 1; \\ \lambda_{i,j}(\lambda(t)) = \beta_{i,j}\lambda_{i+1,2j+1}(t), & t \geq 0; \end{cases} \tag{3}$$

где $0 \leq \beta_{i,j} \leq 1$ – параметры метрики, $i = 0, 1, \ldots$, $j = 0, 1, \ldots, 2^i - 1$.

Поскольку при применении метрики используется только конечное число параметров, то будем считать, что для некоторого l

$$\lambda_{i,j}(t) = 0, \quad \forall i \geq l, \quad j = 0, 1, \ldots, 2^i - 1. \tag{4}$$

В этом случае система рекуррентных уравнений (3) легко решается – последовательно находятся функции $\lambda_{i,j}(t)$ для $i = l-1, l-2, \ldots, 0$.

Отметим, что при таком выборе функции $\lambda(t)$ усредненная по пространству обратная функция к мере шара будет самоподобной для мер Бернулли.

При $\lambda(t) = 0$ получаем метрику, которую будем обозначать через ρ_0.

Подчеркнем, что метрика (1) билипшицево эквивалентна метрике ρ_0 для любой функции $\lambda(t)$, т.е.

$$e^{-1}\rho_0(\boldsymbol{x}, \boldsymbol{y}) \leq \rho(\boldsymbol{x}, \boldsymbol{y}) \leq \rho_0(\boldsymbol{x}, \boldsymbol{y}). \tag{5}$$

Отметим, что, согласно терминологии [1], ρ является *слабой* метрикой (weak или near-metric), поскольку неравенство треугольника для нее выполняется с некоторой константой $C > 1$.

Поскольку каждая точка $\boldsymbol{x}$ имеет бесконечное число координат, то для прикладных вычислений нужно ограничить число координат, используемых при расчетах. Для этого определяется *усечение* метрики, которое использует только m первых координат точки.

Определим *усечение* $\rho^{(m)}$ метрики ρ следующим образом:

$$\rho^{(m)}(\boldsymbol{x}, \boldsymbol{y}) = e^{-\alpha^{(m)}(\boldsymbol{x}, \boldsymbol{y})}, \tag{6}$$

где

$$\begin{aligned} &\alpha^{(0)}(\boldsymbol{x}, \boldsymbol{y}) = 0; \\ &\alpha^{(m)}(a\boldsymbol{x}, b\boldsymbol{y}) = \begin{cases} \alpha^{(m-1)}(\boldsymbol{x}, \boldsymbol{y}), & a = b; \\ \lambda(\alpha^{(m-1)}(\boldsymbol{x}, \boldsymbol{y})), & a \neq b. \end{cases} \end{aligned} \tag{7}$$

Нетрудно видеть, что

1. Для нахождения $\alpha^{(m)}(\boldsymbol{x}, \boldsymbol{y})$ нужны только первые m координат точек $\boldsymbol{x}$, $\boldsymbol{y}$;
2. $\alpha^{(m)}(\boldsymbol{x}, \boldsymbol{y})$ является линейной функцией по каждому параметру $\beta_{i,j}$, $i = 0, 1, \ldots, l-1$, $j = 0, 1, \ldots, 2^i - 1$.

Покажем, что предлагаемый класс метрик достаточно широкий – с любой точностью можно реализовать произвольную функцию $\lambda(t)$.

Утверждение 1. *Для любой строго возрастающей функции $f(x)$, $f(0) = 0$, $f(x) < 1$, и любого $l > 0$ существуют такие значения параметров метрики $\beta_{i,j}$, для которых*

$$\lambda(x_k) = f(x_k), \quad k = 0, 1, \ldots, N = 2^l - 1,$$

где $x_0 = 0 < x_1 < \cdots < x_N$ – некоторый набор точек, $i = 0, 1, \ldots, l-1$, $j = 0, 1, \ldots, 2^i - 1$.

Доказательство. Проведем индукцию по l.

При $l = 1$ имеем

$$\lambda(t) = \begin{cases} \beta_{0,0}, & t \geq 1; \\ 0, & 0 \leq t < 1, \end{cases}$$

поэтому утверждение справедливо при $x_0 = 0$, $x_1 = 1$, $\beta_{0,0} = f(1)$.

Предположим, что утверждение верно для $l - 1$, и покажем, что оно верно для значения l.

Пусть $x'_0 = 0 < x'_1 < \cdots < x'_n$ – набор точек для $l - 1$, $n = 2^{l-1} - 1$.

По предположению индукции существуют такие значения параметров $\beta_{i,j}$, для которых

$$\lambda_{1,0}(x'_k) = \frac{f(x'_k + 1) - f(1)}{1 - f(1)}, \quad \lambda_{1,1}(x'_k) = \frac{f(\lambda(x'_k))}{f(1)}.$$

Положим $x_0 = 0$, $x_{n+1} = 1$ и $\beta_0 = f(1)$. Остальные точки нового набора определим по правилу:

$$x_k = \lambda(x'_k); \quad x_{k+n+1} = x'_k + 1; \quad k = 1, 2, \ldots, n.$$

Подставляя эти значения в (3), получим

$$\lambda(x_k) = \lambda(\lambda(x'_k)) = \beta_0 \lambda_{1,1}(x'_k) = f(x_k),$$
$$\lambda(x_{k+n+1}) = \lambda(x'_k + 1) = (1 - \beta_0)\lambda_{1,0}(x'_k) + \beta_0 = f(x'_k) = f(x_{k+n+1}).$$

□

Ограничение $\lambda(t) < 1$ позволяет организовать быстрый поиск ближайших точек.

Пусть заданы $n + 1$ строк $\boldsymbol{\xi_0}, \ldots, \boldsymbol{\xi_n}$ длины m.

Построим бинарное дерево (trie) T, листьями которого будут заданные строки. Тогда k-я ближайшая строка к заданной находится с трудоемкостью $O(m)$.

Обоснование вытекает из следующего простого свойства рассматриваемых метрик, которое сформулируем в виде утверждения.

Утверждение 2. *В бинарном дереве T k-я ближайшая строка к заданной лежит в первом поддереве на пути из заданной строки в корень, у которого число листьев больше k.*

2. Оценки энтропии

Рассматриваемые меры μ будем считать эргодическими и инвариантными относительно сдвига.

Напомним, что энтропией (на символ) называется величина

$$h = -\lim_{n\to\infty} \frac{1}{n}\mathbf{E}\ln\mu(C_n(\boldsymbol{\xi})), \tag{8}$$

где $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots)$ – случайная точка в Ω, распределенная по мере μ, а через

$$C_s(\boldsymbol{x}) = \{\boldsymbol{y} \in \Omega : y_1 = x_1, \ldots, y_s = x_s\}$$

будем обозначать цилиндры в пространстве Ω.

Отметим, что натуральный логарифм в определении энтропии выбран для удобства (в нижеприведенной оценке (10) не будет постоянного множителя).

Пусть заданы $n+1$ точек $\boldsymbol{\xi_0}, \ldots, \boldsymbol{\xi_n}$ в пространстве Ω.

Кроме точек, будем использовать следующие вспомогательные параметры:

- k – параметр для контроля применимости оценки. Оценки, полученные при различных значениях k, являются оценками одной и той же величины.

- m – параметр, задающий сечение метрики. Поэтому в вычислениях будут использоваться только первые m координат этих точек, которые можно считать строками длины m.

- $l < m$ – параметр, задающий число параметров метрики (1) - (4).

- $0 \le \beta_{i,j} \le 1$ – параметры метрики (1) - (4), $i = 0, 1, \ldots, l-1$, $j = 0, 1, \ldots, 2^i - 1$.

Приведенный в [5] процесс построения оценки состоит из двух шагов.

- Построение вспомогательной статистики

$$r_n^{(k,m)}(\rho) = \frac{1}{n+1}\sum_{j=0}^{n} \max_{i:i\neq j}{}^{(k)} \alpha^{(m)}(\boldsymbol{\xi_i}, \boldsymbol{\xi_j}), \tag{9}$$

  где $\max^{(k)}\{X_1, \ldots, X_N\} = X_k$, если $X_1 \ge X_2 \ge \cdots \ge X_N$.

- Оценка обратной энтропии определяется по формуле

$$\eta_n^{(k,m)}(\rho) = k\left(r_n^{(k,m)}(\rho) - r_n^{(k+1,m)}(\rho)\right). \tag{10}$$

Таким образом, построенная оценка $\eta_n^{(k,m)}(\rho)$ зависит от $2^l - 1$ параметров $\beta_{i,j}$ и является линейной функцией по каждому параметру.

3. Свойства оценки

Кроме эргодичности и инвариантности относительно сдвига, на рассматриваемые меры μ наложим следующее ограничение:

$$\exists a, b > 0 \ : \ \mu(C_n(\boldsymbol{x})) \le be^{-an}, \quad \forall n > 0, \tag{11}$$

для почти всех $\boldsymbol{x} \in \Omega$.

В силу билипшицевоей эквивалентности (5) всех метрик из рассматриваемого класса применимы теорема 1 и утверждение 8 из [5], из которых получаем

Утверждение 3. *Пусть* $\boldsymbol{\xi_0}, \ldots, \boldsymbol{\xi_n}$ *– независимые точки в* Ω*, распределенные по мере* μ *и* $k = O(\log n)$*, тогда для любой метрики* (1) – (4)

$$\lim_{n\to\infty} \frac{\mathbf{E} r_n^{(k,\infty)}(\rho)}{\ln n} = \frac{1}{h}.$$

Утверждение 4. *Пусть выполнено условие* (11)*, тогда существует такая константа* c*, что при* $m > c \ln n$

$$\mathbf{E} r_n^{(k)}(\rho) - \mathbf{E} r_n^{(k,m)}(\rho) = O(n^{-1})$$

для любой метрики (1) – (4)*.*

Здесь и далее $\mathbf{E} r_n^{(k)}(\rho) = \mathbf{E} r_n^{(k,\infty)}(\rho)$.

Доказательство проводится так же, как в [3].

Применяя утверждение 2 и действуя так же, как в [3], получим

Утверждение 5. *Для некоторой константы* C

$$\mathbf{D} r_n^{(k,m)}(\rho) \le C \frac{m^2 k^2}{n}.$$

Итак, для определения точности оценки нужно находить смещение. Смещение проще находить при $m = \infty$.

При доказательстве теоремы 1 из [5] получена формула (4.8), которую приведем в следующем виде.

Утверждение 6. *Пусть* ρ *– произвольная метрика на пространстве* Ω*, тогда*

$$\mathbf{E} r_n^{(k)}(\rho) = \frac{n!}{(k-1)!(n-k)!} \int_0^1 \chi(t)\, t^{k-1} (1-t)^{n-k-1}\, dt. \tag{12}$$

Здесь через $\chi(t)$ *обозначается функция*

$$\chi(t) = \int_\Omega \nu(t, \omega)\, d\mu(\omega), \tag{13}$$

а через $x = \nu(t, \omega)$ *– обобщенная обратная функция к мере шара* $t = \mu(B(\omega, e^{-x}))$ *(при заданном* ω*), т.е.*

$$\nu(t, \omega) \ = \sup\{x \ : \ \mu(B(\omega, e^{-x})) < t\}, \tag{14}$$

$B(\omega, r)$ *– шар радиуса* r *с центром в точке* ω*.*

4. Бернуллиевские меры

Найдем функцию $\chi(t)$ для меры Бернулли и метрики (1) – (4). Как обычно, обозначим

$$p = \mu(C_1(1\omega)), \quad q = \mu(C_1(0\omega)).$$

Для определенности будем считать, что $p \le q$.

Функция $\nu(t,\omega)$ для меры Бернулли и метрики (1) – (4) задается следующими рекуррентными уравнениями:

$$\begin{aligned}
\nu(t,0\omega) &= \begin{cases} \nu\left(\frac{t}{q},\omega\right)+1, & t \le q; \\ \lambda\left(\nu\left(\frac{t-q}{p},\omega\right)\right), & q < t \le 1; \end{cases} \\
\nu(t,1\omega) &= \begin{cases} \nu\left(\frac{t}{p},\omega\right)+1, & t \le p; \\ \lambda\left(\nu\left(\frac{t-p}{q},\omega\right)\right), & p < t \le 1. \end{cases}
\end{aligned} \tag{15}$$

Введем функции $\chi_{i,j}(t)$, положив

$$\chi_{i,j}(t) = \int_\Omega \lambda_{i,j}\left(\nu(t,\omega)\right)\, d\mu(\omega). \tag{16}$$

Перепишем (13) как

$$\chi(t) = p\int_\Omega \nu(t,1\omega)\, d\mu(\omega) + q\int_\Omega \nu(t,0\omega)\, d\mu(\omega).$$

Подставляя (15), получим

$$\chi(t) = \begin{cases} p\chi\left(\frac{t}{p}\right) + q\chi\left(\frac{t}{q}\right)+1, & 0 < t \le p; \\ p\chi_{0,0}\left(\frac{t-p}{q}\right) + q\chi\left(\frac{t}{q}\right)+q, & p < t \le q; \\ p\chi_{0,0}\left(\frac{t-p}{q}\right) + q\chi_{0,0}\left(\frac{t-q}{p}\right), & q < t \le 1. \end{cases}$$

Будем считать, что $\chi(t) = \chi_{i,j}(t) = 1$ при $t \le 0$ и $\chi(t) = \chi_{i,j}(t) = 0$ при $t > 0$, тогда полученное уравнение можно переписать в следующем виде:

$$\chi(t) = p\chi\left(\frac{t}{p}\right) + q\chi\left(\frac{t}{q}\right) + p\chi_{0,0}\left(\frac{t-p}{q}\right) + q\chi_{0,0}\left(\frac{t-q}{p}\right). \tag{17}$$

Аналогичным образом для функций $\chi_{i,j}(t)$ из (3) получаем

$$\begin{aligned}
\chi_{k-1,j}(t) = (1-\beta_{k,j})&\left(p\chi_{k,2j}\left(\frac{t}{p}\right) + q\chi_{k,2j}\left(\frac{t}{q}\right)\right) + \\
&+ \beta_{k,j}\left(p\chi_{k,2j+1}\left(\frac{t-p}{q}\right) + q\chi_{k,2j+1}\left(\frac{t-q}{p}\right)\right), \\
&i = 0,1,\ldots,l; \quad j = 0,1,\ldots,2^i-1.
\end{aligned} \tag{18}$$

Покажем, что параметры метрики можно выбрать так, чтобы смещение было произвольно малым. Для этого достаточно показать, что решением системы (18) может быть любая убывающая функция $\chi_{0,0}(t)$.

В настоящей работе приведем доказательство только для симметричной меры ($p = q = 1/2$). Общий случай доказывается аналогично, но очень громоздко.

Утверждение 7. *Для любой строго убывающей функции $f(x)$, $f(0) = 1$, $f(1) = 0$, и любого $l > 0$ существуют такие значения параметров метрики $\beta_{i,j}$, для которых*

$$\chi_{0,0}(k2^{-l}) = f(k2^{-l}), \quad k = 0, 1, \ldots, N = 2^l - 1.$$

Доказательство. При $p = q = 1/2$ система (18) существенно упрощается.

$$\chi_{k-1,j}(t) = (1 - \beta_{k,j})\chi_{k,2j}(2t) + \beta_{k,j}\chi_{k,2j+1}(2t - 1), \quad i = 0, 1, \ldots, l; \quad j = 0, 1, \ldots, 2^i - 1. \quad (19)$$

Проведем индукцию по l.

При $l = 1$ имеем

$$\chi_{0,0}(t) = \begin{cases} 1, & t \geq 0; \\ \beta_{0,0}, & 0 < t \geq 1/2; \\ 0, & 1/2 < t; \end{cases}$$

поэтому утверждение справедливо при $\beta_{0,0} = f(1/2)$.

Предположим, что утверждение верно для $l - 1$, и покажем, что оно верно для значения l.

Положим $\beta_{0,0} = f(1/2)$.

По предположению индукции существуют такие значения параметров $\beta_{i,j}$, для которых ($i + j > 0$)

$$\chi_{1,0}(k2^{-l+1}) = f_0(k2^{-l+1}), \quad \chi_{1,1}(k2^{-l+1}) = f_1(k2^{-l+1}), \ k = 0, 1, \ldots, 2^{l-1} - 1;$$

где $f_0(x), f_1(x)$ – произвольные убывающие функции.

Положим

$$f_0(x) = \frac{f(x/2) - \beta_{0,0}}{1 - \beta_{0,0}} \quad f_1(x) = \frac{f(x/2 + 1/2)}{\beta_{0,0}}.$$

Подставляя эти функции в (19), получим

$$\chi_{0,0}(k2^{-l}) = (1 - \beta_{0,0})\chi_{1,0}(k2^{-l+1}) + \beta_{0,0} = (1 - \beta_{0,0})f_0(k2^{-l+1}) + \beta_{0,0} = f(k2^{-l}),$$
$$\chi_{0,0}(k2^{-l} + 1/2) = \beta_{0,0}\chi_{1,1}(k2^{-l+1}) = \beta_{0,0}f_1(k2^{-l+1}) = f(k2^{-l} + 1/2),$$
$$k = 0, 1, \ldots, 2^{l-1} - 1.$$

□

Список литературы

1. *Deza M., Deza T.* Encyclopedia of Distances. Springer, 2009.

2. *Grassberger P.* Estimating the information content of symbol sequences and efficient codes // *IEEE Trans. Inform. Theory.* 1989. V. 35. P. 669–675.

3. *Kaltchenko A., Timofeeva N.* Entropy Estimators with Almost Sure Convergence and an $O(n^{-1})$ Variance // Advances in Mathematics of Communications. 2008. V. 2, №1. P. 1–13.

4. *Kaltchenko A., Timofeeva N.,* Rate of convergence of the nearest neighbor entropy estimator // AEU – International Journal of Electronics and Communications. 2010. **64**, №1. P. 75–79.

5. *Timofeev E.A.* Statistical Estimation of measure invariants // St. Petersburg Math. J. 2006. **17**, №3. P. 527–551.

6. *Timofeev E.A.* Bias of a nonparametric entropy estimator for Markov measures // Journal of Mathematical Sciences. 2011. **176**, №2. P. 255–269.

Unbiased Entropy Estimator for Binary Sequences

Timofeev E.A.

P.G. Demidov Yaroslavl State University,
Sovetskaya str., 14, Yaroslavl, 150000, Russia

Keywords: entropy, nonparametric statistic, ball, Bernoulli's measure

A new class of metrics on a space of right-sided infinite sequences drawn from a binary alphabet was introduced.

Сведения об авторе:
Тимофеев Евгений Александрович,
Ярославский государственный университет им. П.Г. Демидова,
д-р физ.-мат. наук, профессор кафедры теоретической информатики